Magyar mondatok SVM alapú szintaxiselemzése

Iván Szilárd[1], Ormándi Róbert[2], Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
szilivan@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{ormandi,kocsor}@inf.u-szeged.hu

Kivonat: A nyelvtechnológiai alkalmazások egyik fontos elemzése a szintaxis-elemzés. Bemutatásra kerül egy gépi tanuláson alapuló szintaxis elemző, mely az SVM alapú megközelítést alkalmazza. A használt algoritmusok elméleti és implementációs részleteinek bemutatásán túl, átfogó teszteléssel igazoljuk a módszer alkalmazhatóságát. A módszer további érdekessége, hogy a strukturált kimenetű tanulás paradigmáját követi.

1 Bevezetés

A szintaxis elemzés a természetes nyelvi feldolgozás elemzéseinek azon csoportja, melyeknek célja a mondatok nyelvtani struktúrájának felderítése. Ez a struktúra leggyakrabban egy hierarchikus szerkezet, ahol a legnagyobb egység a mondat, legkisebb egységei pedig az alapszimbólumok (például a szavak szófajai, vagy azok POS kódjai). Az ilyen típusú szerkezetet általában egy fával, az úgynevezett szintaxis fával szokás reprezentálni, melynek gyökerében az egész mondatnak megfelelő csúcs áll, levelei az alapszimbólumokat, a belső csúcsok pedig az egyéb nyelvtani egységeket reprezentálják.

Az ilyen típusú elemzésnek rendkívül fontos szerepe van a természetes nyelvi feldolgozás számos területén, hiszen egy mondat szintaxisfájának helyes meghatározása alapvető fontosságú a magasabb szintű szövegfeldolgozáshoz (például szemantikai elemzés, vagy gépi fordítás).

Szintaxis elemzésre alapvetően kétféle megközelítés létezik, az egyik a szakértők által megadott összefüggéseken alapuló, a másik a gépi tanulást előtérbe helyező eljárások. Manapság a figyelem az utóbbi módszerekre összpontosul, angol nyelvre igen hatékony algoritmusok kerültek kidolgozásra, de a magyar nyelv sajátosságai (nyelvi variabilitás) miatt ezek változatlan formában történő alkalmazása jelentős hatékonyságveszteséggel jár [1], [2].

2 SVM alapú megközelítés

Jelen publikációban bemutatunk egy gépi tanuláson alapuló szintaxis-elemző eljárást, amely a manapság intenzíven kutatott SVM alapú megközelítést követi [5]. A kidol-

gozott eljárás a szintaxisfákat mint, adott valószínűségi környezet független nyelvtan feletti derivációs fákat értelmezi. Ezeket a fákat jól jellemzi, hogy a deriváció során az egyes szabályok hányszor lettek alkalmazva. A megközelítés lényege, hogy a szabályok alkalmazásának eloszlásának becslését végzi [3], [4]. Az itt előálló feladat, átalakítható olyan formára, mely a manapság intenzíven kutatott margó maximalizáló eljárások segítségével oldható meg. A módszer algoritmikus részleteinek bemutatásán túl, egy releváns gyakorlati feladaton keresztül igazoljuk a bevezetett eljárás létjogosultságát.

3 Korpusz

Az algoritmus teszteléséhez szükséges mondatok, és hozzájuk tartozó szintaxis-fák a Szeged Korpusz adattárából származnak. A korpusz több témakörben (iskolai, jogi, számítógépes, szépirodalmi, üzleti) tartalmaz szövegeket, amelyeken nyelvészek a különféle elemzéseket, mint morfológiai elemzés, szófaji egyértelműsítés, szintaxis elemzés.
A tanításhoz és teszteléshez használt mondatok az üzleti témakörben található mondatokból kerültek ki. A tanulás-tesztelés során használt szintaxisfák teljesen általános struktúrával rendelkeznek.

4 Eredményeink

Annak mérése, hogy az elemzés eredményeképpen előálló fa, mennyire jó, azaz mennyire hasonlít az elvárt szintaxisfához nem könnyű feladat. A szakirodalomban, erre három elterjedt mértéket szoktak használni:

- *Pontosság (precision):* a helyesen felismert szócsoportok számának és az összes felismert szócsoport számának hányadosa.
- *Fedés (recall):* a helyesen felismert szócsoportok számának és a mintában ténylegesen szereplő szócsoportok számának hányadosa.
- *F1-mérték:* 2*Pontosság*Fedés / (Pontosság+Fedés), azaz a Pontosság és a Fedés harmonikus átlaga.

Látható, hogy mindhárom mérték 0 és 1 között mozog, és minél nagyobb értéket vesz fel, annál ,,jobbnak" mondható az eredmény. Az 1. táblázat összefoglalja a mérési eredményeink. A táblázatban szereplő értékek F1-mértékben értendőek.

1. Táblázat: Eredmények a használt mondatok hosszának függvényében.

Mondat hossza	Tanító adatbázis	Teszt adatbázis
15-20	87.7%	89.2%
21-25	87.2%	89.0%
26-30	86.1%	86.5%
31-35	85.1%	83.3%
36-40	84.3%	79.6%
41-45	84.0%	78.7%
3-45	86.8%	87.0%

Bibliográfia

1. Brill, E.: Transformation-Based Learning. PhD thesis, University of Pennsylvania, (1993)
2. Hócza, A.: Teljes mondatszintaxis tanulása és felismerése. MSZNY (2004) 127-135
3. Joachims, T.: A support vector method for multivariate performance measures. Twenty-Second International Conference on Machine Learning (2005)
4. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables. Journal of Machine Learning Research (JMLR) (2005) 1453-1484
5. Vapnik, V.: Statistical learning theory. Wiley and Sons Inc (1998)